DOCUMENT RESUME

ED 443 836 TM 031 481

AUTHOR	Manalo, Jonathan R.; Wolfe, Edward W.
TITLE	The Impact of Composition Medium on Essay Raters in Foreign Language Testing.
PUB DATE	2000-04-00
NOTE	14p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000). Funding provided by the TOEFL program.
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Computer Assisted Testing; *Essay Tests; *Evaluators; *Handwriting; Interrater Reliability; Language Tests; Test Format; *Validity; Word Processing
IDENTIFIERS	Paper and Pencil Tests; *Test of English as a Foreign Language

ABSTRACT

Recently, the Test of English as a Foreign Language (TOEFL) changed by including a writing section that gives the examinee an option between computer and handwritten formats to compose their responses. Unfortunately, this may introduce several potential sources of error that might reduce the reliability and validity of the scores. The seriousness of these sources of error was studied by examining the quality of the ratings and the distribution of construct-irrelevant variance. Participants were 152,951 TOEFL examinees who participated in regular TOEFL administrations. Results indicate that raters have slightly better agreement for word-processed essays than handwritten essays. The generalizability analyses suggest that this is caused by the interaction of examinees x items x readers confounded with undifferentiated error. (Author/SLD)

Running head: COMPOSITION MEDIUM IMPACT

The Impact of Composition Medium on Essay Raters in Foreign Language Testing

Jonathan R. Manalo

Edward W. Wolfe

Michigan State University

Abstract

Recently, the Test of English as a Foreign Language changed by including a writing section that gives the examinee an option between computer and handwritten formats to compose their responses. Unfortunately, this may introduce several potential sources of error that might reduce the reliability and validity of the scores. We attempted to identify the seriousness of this by examining the quality of the ratings and the distribution of construct-irrelevant variance. Results indicated raters have slightly better agreement for word-processed essays than handwritten essays. The generalizability analyses suggests this is caused by the interaction of examinees × items × readers confounded with undifferentiated error.

The Impact of Composition Medium on Essay Raters in Foreign Language Testing

Currently, the Test of English as a Foreign Language (TOEFL) requires examinees to complete a computer-based multiple-choice section as well as a writing section administered by either computer or pencil and paper. However, by administering the writing section in both media, the potential for construct-irrelevant variance to negatively impact the scores becomes a greater concern. That is, there is a higher likelihood that the composition medium may influence the reliability of the scores and the accuracy of the inferences that are based on those scores. A potential reason for this error may be differential rater perceptions. Raters may perceive an essay composed in one medium to be qualitatively different from an essay composed in another medium. For example, Powers. Fowles, Farnum, and Ramsey (1994) found that when an essay composed in handwriting is subsequently transcribed to a typewritten format, raters prefer to read the handwritten essays. Similarly, raters may have higher expectations for the computer-based essays (Arnold, Legas, Obler, Pacheco, Russell, & Umbdenstock, 1990). Other researchers have found that raters potentially perceived differences in the apparent length, ease of reading, and surface correctness of handwritten and word-processed essays (Bridwell, Sirc, & Brooke, 1985; Hawisher, 1987; Levin, Riel, Rowe, & Boruta, 1985). Fortunately, raters can be trained to compensate for their perceptual differences (Powers et al., 1994). Hence, a specific rater's judgements can be made to be consistent across the different formats.

Research Questions

Research, however, has yet to provide evidence of this for foreign language examinees. Our purpose is to determine the degree rater judgements are affected by computer-based and handwriting composition mediums for a foreign language examination, specifically, the TOEFL.

Our research question is: "To what degree are rater judgements affected by computer-based and handwriting composition mediums for the TOEFL?"

Method

Examinees

Participants were 152,951 TOEFL examinees (males = 53.5%, females = 46.5%) from 223 countries that participated in a regular administration between 1/24/98 and 2/9/99 of the computer-based TOEFL test. Most of the examinees indicated taking the test for admittance into undergraduate (39%) or graduate studies (45%). The others indicated non-academic reasons (16%). Each examinee was administered a computer-based multiple-choice section, while being offered the choice of word processor (51.5%) or handwriting (48.5%) for the single prompt of the writing section.

TOEFL

The TOEFL test consists of four sections: (a) listening, (b) ,tructure, (c) reading, and (d) writing. As stated previously, the examinee is administered the computer format for the multiple-choice sections listening, structure, and reading, while for the writing section, the examinee is given the choice between paper and pencil or a word processor. Our analyses focused only on the responses to the writing section.

The writing section measures the examinee's proficiency in writing English. A single prompt requires examinees to compose a response that accomplishes the following: (a) generate, organize, and develop ideas; and (b) support those ideas using examples or evidence. Examinees are given 30 minutes to respond to the prompt. Operationally, the writing section is scored on a scale ranging from 1 to 6 by two trained and independent readers (i.e., trained in interpreting the response to TOEFL standards, to score across multiple topics, and to use the software which essays are distributed to the readers and scores are recorded; and independent from knowing the score assigned by the other reader). The average of the two scores (i.e., writing composite score) is assigned to the essay unless there is a discrepancy between the two readers. In that case, a third independent reader rates the essay, and the score is resolved (ETS, 1999).

Analysis

Quality of the Ratings. The analysis of the quality of the ratings was accomplished by examining scores assigned to handwritten and word-processed essays separately using several reliability indices. The first index, the Pearson product moment correlation (r), was computed using the scores assigned by the two randomly selected readers for each examinee's essay. The second index, or group of indices, is the proportion of perfect, adjacent, and outside-of-adjacent agreement between the two reader scores. These were determined by the absolute differences between the two scores that were equal to 0, 1, and 2 or more points on the 6-point TOEFL rating scale, respectively. The third index is Cohen's coefficient kappa (κ)

$$\kappa = \frac{P_o - P_e}{1 - P_e}.$$

This coefficient indicates the degree to which readers agree beyond the level expected by chance, where P_0 is the observed proportion of ratings in perfect agreement and P_e is the expected proportion of ratings in perfect agreement based on the marginal distributions of ratings assigned by the two readers.

Generalizability Analyses. The generalizability analyses examines the scores assigned to handwritten and word-processed essays separately to determine the extent ratings are influenced by construct-irrelevant variance. Here, we used a $(e : i) \times r$ design, where examinees (e) are nested within items (i) and crossed with readers (r). Variance components and dependability coefficients (ϕ) were estimated between the two modes of composition and then compared. The ϕ coefficients were calculated using

$$\phi = \frac{\sigma^2_{(e:i)}}{\sigma^2_{(e:i)\cdot r} + \frac{\sigma^2_i}{n_i} + \frac{\sigma^2_{(e:i)\cdot r,e}}{n_i n_r}}.$$

where σ^2 is the variance component for a facet of the measurement design (i.e., a main effect for examinees, e, items, i, or raters, r, or some interaction of these sources of variance), and n is the number of elements of that facet across which observed scores are recorded.

Unfortunately, because the sample size was very large (152,951), estimating variance components for all the data in a single estimation 'run' was limited by the software capabilities of Proc Varcomp in SAS (1998). However, by dividing the data in each medium into five mutually exclusive data sets, estimation of the variance components was possible. That is, we averaged the variance components (weighted by the sample size of the respective data set) of the five mutually exclusive data sets for each medium. The ϕ coefficients were estimated using these weighted averages.

Results

Quality of the Ratings

The indices of rating quality are shown in Table 1. Specifically, for each composition medium, we show the following indices: (a) the Pearson product moment correlation (r), (b) Cohen's coefficient kappa (κ), and (c) the percent of perfect, adjacent, and outside-of-adjacent agreement. The table reveals that, in general, the differences in the quality of the ratings, although only modest, are in favor of the word-processed essay scores. That is, it was generally easier for readers to agree on scores for the word-processed essays than for the handwritten essays. This was true regardless of the index we considered. Specifically, r is .67 for readers of handwritten essays and .74 for readers of word-processed essays. κ is .30 for readers of handwriting essays and .35 for readers of word-processed essays. The agreement indices are roughly equal, but slightly in favor of readers of computer based essays. Specifically, the degree of perfect agreement (a difference of zero) is 50% for readers of handwritten essays and 51% for word-processed essays. The degree of adjacent agreement (an absolute difference of one) is 45% for handwritten essays and 44% for word-processed essays. The degree of outside agreement (an absolute difference of two or more) is 6% for readers of handwritten essays and 5% for readers of word-processed essays.

Generalizability Analyses

The generalizability analyses results are shown in Table 2. The analysis reveals that the scores for word-processed essays are more reliable than scores for handwritten essays. Specifically, the ϕ coefficient is .79 for handwriting and .85 for word processing. The variance components reveal that the variance that is accounted for by the two mediums is equal or approximately equal between each mode for the item, reader, and item × reader facets. It also suggests that the differences in random error across modes can be attributed to the undifferentiated error. In other words, there is more variance attributed to true variance (i.e., examinees) in word processing (73%) than in handwriting (65%), while there is less variance attributed to undifferentiated error variance in word-processing (23%) than handwriting (30%).

Discussion

To summarize, our findings revealed the following: (1) r, κ, the degree of perfect agreement, and the ϕ coefficient were lower when raters evaluated handwritten essays than when they evaluated word-processed essays; (2) the degree of non-perfect agreement (adjacent and outside) was lower for readers of word-processed essays than for readers of handwritten essays; and (3) more variance is attributed to true variance for word processors than handwritten essays, while more error variance is attributed to undifferentiated error variance for handwritten essays than word-processed essays.

Although these differences are small, specifically, the reliability differences are no greater than .07, and the degree of agreement (or non-agreement) does not differ by more than 7%, these findings are consistent with other research (e.g., Bridgeman & Cooper, 1998). Given the high stakes of large-scale examinations, these small but consistent findings take on greater meaning. Our findings suggest that having different writing formats in large-scale examinations, at least in the case of the TOEFL, may introduce construct-irrelevant variance into the rater's judgements. Our generalizability analyses suggests that this is possibly the result of the interaction of examinees × items × readers in conjunction with the undifferentiated error variance. In other words, the cause of the poorer agreement among scores for handwritten essays may lie in the non-uniformity of these essays (i.e., differences in handwriting quality, etc.)—a conclusion that was also drawn by Bridgeman and Cooper (1998). Therefore, our findings suggest that the scores for handwritten essays will have decreased reliability, and consequently, the inferences that are based on those scores may have decreased validity. Hence, the fairness of using such scores to make decisions about these individuals is questionable.

There are three general limitations to our study. First, the results of the reliability analyses are not as convincing or informative as they could be if a better design or more appropriate analytic method was employed. For example, our design did not allow us to examine sources of error that would have been informative, such as the examinee × rater interaction.

Here, this was confounded with undifferentiated error. Second, examinee ability confounds any difference between media (i.e., selection is a threat to internal validity). It may well be that the observed differences are due to differences in the variance of the two populations. Third, computer hardware and software limitations prevented us from performing the cleanest and most sophisticated analyses that we could have performed. For example, we may have been able to more directly estimate the variance components for the generalizability study and to use better methods for combining variance component estimates (i.e., weighting each variance component by its estimation variance rather than by its sample size) from the subsets of data that we analyzed separately if we were to have used Proc Mixed in SAS rather than Proc Varcomp. Future research could perform more complete and more sophisticated generalizability analyses of these data.

References

Arnold, V., Legas, J., Obler, S., Pacheco, M.A., Russell, C., & Umbdenstock, L. (1990). Direct writing assessment: A study of bias in scoring hand-written vs. word-processed papers (unpublished paper). Whittier, CA: Rio Hondo College.

Bridgeman, B., & Cooper, P. (1998). Comparability of scores on word-processed and handwritten essays on the graduate management admissions test. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Bridwell, L., Sirc, G., & Brooke, R. (1985). Case studies of student writers. In S.W. Freedman (Ed.), The acquisition of written language (pp. 172-194). Norwood, NJ: Ablex.

ETS. (1999). Description of the computer-based TOEFL test. [Internet Web Page] Available: http://www.teofl.org/descbcbt.html. Princeton, NJ: Author.

Hawisher, G. (1987). The effects of word processing on the revision strategies of college freshmen. Research in the Teaching of English, 21, 145-159.

Levin, J., Riel, M., Rowe, R., & Boruta, M. (1985). Muktuk meets jacuzzi: Computer networks and elementary school workers. In S.W. Freedman (Ed.), The acquisition of written language (pp. 160-171). Norwood, NJ: Ablex.

Powers, D.E., Fowles, M.E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. Journal of Educational Measurement, 31, 220-233.

SAS. (1998). Statistical Analysis Software. Cary: SAS Institute.

Table 1

Indices of Rating Quality for Readers of Handwriting and Word-Processing Composition Mediums of the TOEFL

	Composition Medium	
	Handwriting	Word Processing
r	.67	.74
κ	.30	.35
% Perfect	50%	51%
% Adjacent	45%	44%
% Outside	6%	5%

Note. r = Pearson correlation coefficient. κ = Cohen's coefficient kappa. % Perfect, % Adjacent, and % Outside are the proportion of absolute differences between the two reader scores that were equal to 0, 1, and 2 or more points on the 6-point TOEFL rating scale, respectively. Sample size is 152,951.

Table 2

Generalizability Analyses for Readers of Handwriting and Word-Processing Composition Mediums of the TOEFL

	Composition Medium	
	Handwriting	Word Processing
ϕ	.79	.85
Variance Component		
Examinee	65%	73%
Item	0.3%	0.3%
Reader	1.7%	1.6%
Item × Reader	2.5%	1.7%
Undifferentiated	30%	23%

Note. ϕ is the dependability coefficient. Sample size is 152,951.

Acknowledgements

This project greatly benefited from the input of our colleagues. Specifically, Claudia Gentile provided input into the design and data collection for this study. Pat Carey, Robbie Kantor, Yong-Won Lee, Philip Oltman, and Ken Sheppard each provided guidance in obtaining and interpreting the data. In addition, James Algina, Robert Brennan, Chris Chiu, and David Miller gave us advice during various stages of data analysis. We also thank the TOEFL program for providing us with the funds to carry out this work.